Astronomical Data Analysis with Commodity Components

Michael S. Warren, T-6; John Wofford, Columbia Univ.

During the next decade, large astronomical observing projects will generate more than 1000 times as much observational data as has been gathered in all of our history. Storage, analysis, and management of this information will require significant advances in computing technology. As an initial step in this process, we have developed and deployed an astronomical data system based on open-source software and commodity hardware with a storage capacity of 112 terabytes (TB) in immediately accessible disk arrays for a total cost of $95,127. This approach is scalable to a petabyte of storage for less than $1M. In the same way in which special-purpose telescopes are now required to obtain the best catalogs of objects in the sky, a focused effort involving state-of-the-art parallel computer hardware and software is required to analyze this data and model the evolution of the Universe, which led to the observed distribution of stars and galaxies.

Persistent data storage is a fundamental prerequisite for all information science and technology projects. The advent of commodity microprocessors with adequate floating-point performance and low-priced fast ethernet switches contributed to the emergence of Beowulf clusters in the mid-1990s, which revolutionized parallel computing at the departmental scale [1-3,4]. We are currently in the midst of a similar revolution in scalable data storage, due to the dramatic decline in the price of commodity disk drives and 10-gigabit (Gb) networking technologies.

Vast amounts of relatively inexpensive storage offer significant opportunities to develop new approaches to scientific problems. The cost for SATA disk storage is currently about $0.35 per gigabyte for 1-TB drives. A single fault-tolerant RAID-6 node can store 14 TB in 3U of rack space for a total cost of about $10 K (see Table 1). Used in a parallel cluster environment, a petabyte disk array with achievable read/write bandwidth that greatly exceeds available local and wide-area networking technology is possible.

The problem we address here is optimizing the price/capacity of information storage at the 100-terabyte-to-petabyte scale, while maintaining acceptable application performance within the domain of astronomical data processing applications. Implied in this price/capacity goal is minimizing overall costs over the life-cycle of the system, including administration and maintenance. Acceptable application performance also includes minimizing the possibility of data loss and downtime due to hardware failures, while realizing that there are trade-offs in complexity and cost vs reliability.

The optimal way to move large files was with the *mpscp* software from Sandia National Laboratory, which, using four simultaneous TCP streams, obtained transfer rates of 380 MB/s between two nodes of the cluster, which is near the maximal rate at which a node can write data. We performed a CRC checksum on the 3.5 TB of Sloan Digital Sky Survey image data, consisting of 1.6 million files, in 2,364/s, for an overall processing rate of 686 files per/s, with an aggregate read bandwidth of 1.5 GB/s. Running *sextractor* to extract astronomical objects from each image file completed in 28686/s, for a processing rate of 56 images per/s, or equivalently, 172 megapixels/s. Performing a plane-to-plane reprojection using *mProjectPP* from Montage processed 112,000 files in 54,700/s, for an overall rate of 2 images per/s. An example reprojected image is shown in Fig. 2.

Fig. 1. The deployed storage system, consisting of eight 3U storage nodes and an 8-port 10-gigE switch. The usable storage capacity of the system is 112 TB. The total cost of the cluster was $95 K.

We executed a 32-processor parallel job on the eight nodes of the cluster using our N-body code with a test data set of 400 million particles. The actual write bandwidth to disk during the data dump was 2.6 GB/s. Restarting the code from a data dump required reading in the 12.8-GB/s file, which resulted in a read bandwidth of 1.8 GB/s. The overall floating point performance obtained by the cluster executing the N-body code was 97.8 Gflops. We created an animation

of the evolution of the dark matter in the Universe, which required reading in every data dump produced by a simulation, and projecting the 3D mass density into a 2D image. For this analysis task, the cluster supported reading each 12.8 GB file in 7.5/s, for a read bandwidth of 1.7 GB/s. An example image from this process is shown in Fig. 3.

Reliability of the system has been excellent, with no disk failures in eight weeks of operation. This result is consistent with Google's recent statistics on using over 100,000 disk drives [2].

Our experience with the described hardware configuration has been positive. It provides a reasonable environment to work within, while providing immediately accessible disk storage at a cost that is almost within a factor of two of the cost of bare disk drives. The software environment should be familiar to any Linux user, and does not require any specialized system administration skills. In summary, we have demonstrated a flexible and usable storage system capable of storing 112 TB of information at a cost of $95,127, or a price/capacity of $850 per terabyte.

This project won the Storage Challenge Award at the annual SC2008 conference on high-performance computing, networking, and storage in Reno, NV.

For more information contact Michael S. Warren at msw@lanl.gov.

[1] D.J. Becker, et al., In Proceedings of the 1995 *International Conference on Parallel Processing (ICPP)*, 11-14 (1995).
[2] E. Pinheiro, W.D. Weber, and L.A. Barroso, "Failure trends in large disk drive population," Proceedings of the *5th USENIX Conference on File and Storage Technologies (FAST07)* (2007).
[3] M.S. Warren, et al., Proceedings of the *International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'97)*, 1372-1381 (1997).
[4] M.S. Warren et al., "Pentium Pro inside: I. A teecode at 430 Gigaflops on ASCI Red, II. Price/performance of 50 Mflop on Loki and Hyglac," *Supercomputing '97*, Los Alamitos, IEEE Comp. Soc. (1997).

Funding Acknowledgments
- National Aeronautics and Space Administration

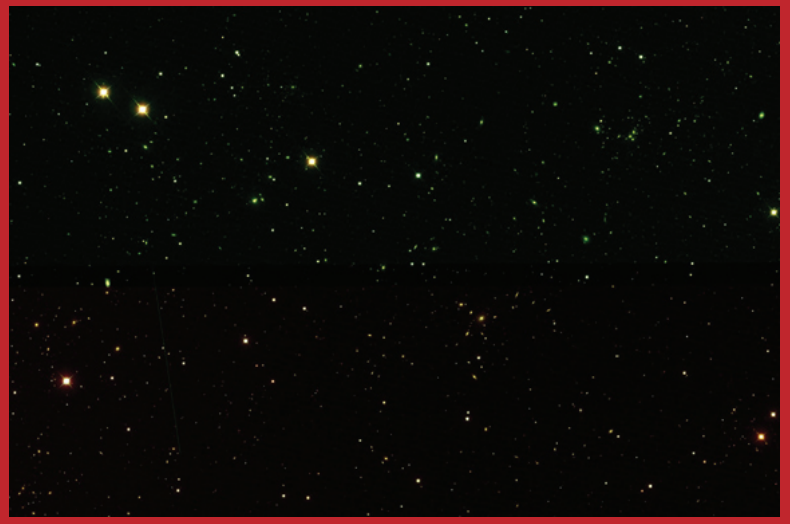

Fig. 2. An example image from the Sloan Digital Sky Survey that was reduced and reprojected using the system described here.

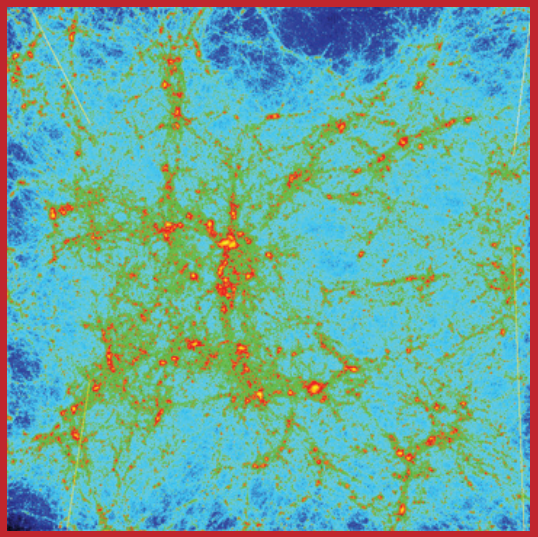

Fig. 3. An image representing the distribution of dark matter in the universe, which was produced from a 128 million particle N-body simulation using the system described here.